# Beyond Frame-level CNN: Saliency-Aware 3-D CNN With LSTM for Video Action Recognition

Xuanhan Wang, Lianli Gao, Jingkuan Song, *Member, IEEE*, and Hengtao Shen, *Senior Member, IEEE*

***Abstract*—Human activity recognition in videos with convolutional neural network (CNN) features has received increasing attention in multimedia understanding. Taking videos as a sequence of frames, a new record was recently set on several benchmark datasets by feeding frame-level CNN sequence features to long short-term memory (LSTM) model for video activity recognition. This recurrent model-based visual recognition pipeline is a natural choice for perceptual problems with time-varying visual input or sequential outputs. However, the above-mentioned pipeline takes frame-level CNN sequence features as input for LSTM, which may fail to capture the rich motion information from adjacent frames or maybe multiple clips. Furthermore, an activity is conducted by a subject or multiple subjects. It is important to consider attention that allows for salient features, instead of mapping an entire frame into a static representation. To tackle these issues, we propose a novel pipeline, saliency-aware three-dimensional (3-D) CNN with LSTM, for video action recognition by integrating LSTM with salient-aware deep 3-D CNN features on videos shots. Specifically, we first apply saliency-aware methods to generate saliency-aware videos. Then, we design an end-to-end pipeline by integrating 3-D CNN with LSTM, followed by a time series pooling layer and a softmax layer to predict the activities. Noticeably, we set a new record on two benchmark datasets, i.e., UCF101 with 13 320 videos and HMDB-51 with 6766 videos. Our method outperforms the state-of-the-art end-to-end methods of action recognition by 3.8% and 3.2%, respectively on above two datasets.**

***Index Terms*—Action recognition, deep learning, three-dimensional (3-D) convolution, LSTM, saliency-aware.**

## I. Introduction

RECOGNIZING human actions in videos has received a significant amount of attention in the research communities [1], [11], [19], [21], [23]. Different types of action recognition algorithms have been recently introduced. In this letter, we divide the video human action recognition into two categories: 1) local space-time feature designing [31]–[33]; and 2) deep learning based techniques.

Human action recognition can be viewed as a pattern recognition problem and hand-crafted features used in image processing [3], [27], [28] have been successfully used for image recognition. Therefore, previous work has directly extended the image action algorithms with hand-crafted spatial and temporal features for recognizing human action in videos. The spatial and temporal features are used to characterize visual appearance and motion dynamics, respectively [1], [11], [23]. For example, Dollar *et al.* [1] presented a behavior recognition algorithm based on spatio-temporal features, which are extracted via anchoring off the direct three-dimensional (3-D) and two-dimensional (2-D) interest points from spatio-temporally windowed data. Wang and Schmid [23] proposed the improved dense trajectories (iDT) by explicitly estimating camera motion. To date, many classical image features used in computer vision [4], [5], [17], [26], [29], [30] have been generalized to video action recognition, such as 3D-SIFT [15], extended SURF, HOG3D, spatio-temporal feature [1], motion boundary histograms (MBH), histograms of optical flow (HOF), and iDT [23]. Among them, MBH has been shown to perform better than HOF for the reason that MBH is robust to camera motion. And iDT give the best results, which performs better than the combination of HOF + MBH. In general, these methods give good results in some challenging datasets (UCF-101 [18], HMDB-51 [10]) by encoding previous local spatial-time features into high-dimensional space. However, their performance often degrades when being applied to more realistic and complex video settings due to the large variations within action categories and other video issues [19], [21].

In order to improve the performance of action recognition, some recent efforts have been proposed to directly apply deep learning models to learn video representation for video action recognition and promising results were obtained [8], [9], [16], [25]. Compared with image action recognition, human actions in video sequences are 3-D signals consisting of visual appearance that dynamically evolves over time [19]. As a result, there are some attempts to change 2-D convolutional neural networks (CNN) or utilize other deep network modules for encoding actions' temporal information. For instance, Ji *et al.* [7] extend the 2-D CNN model into 3-D domain. The proposed 3-D CNN model extracts features from both spatial and temporal dimensions by performing 3-D convolutions, thereby capturing the motion information encoded in multiple adjacent frames. Simonyan *et al.* [16] aimed to capture the complementary information on appearance from still frames and motion between adjacent frames. As a result, they proposed a two-stream ConvNet architecture that incorporates spatial and temporal networks.

Manuscript received July 6, 2016; revised September 7, 2016; accepted September 7, 2016. Date of publication September 20, 2016; date of current version March 14, 2017. This work was supported in part by the Fundamental Research Funds for the Central Universities under Grant ZYGX2014J063 and in part by the National Natural Science Foundation of China under Grant 61502080. The associate editor coordinating the review of this manuscript and approving it for publication was Prof. J.-M. Guo. *(Corresponding author: Lianli Gao.)*

X. Wang is with the University of Electronic Science and Technology of China, Chengdu 610051, China (e-mail: wangxuanhan@std.uestc.edu.cn).

J. Song is with the University of Trento, Trento 38122, Italy (e-mail: jingkuan.song@unitn.it).

H. Shen is with the School of Information Technology and Electrical Engineering, the University of Queensland, Brisbane, Qld. 4072, Australia (e-mail: shenht@itee.uq.edu.au).

L. Gao is with the School of Computer Science and Engineering, University of Electronic Science and Technology of China, Chengdu 611731 (e-mail: lianli.gao@uestc.edu.cn).

Color versions of one or more of the figures in this letter are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/LSP.2016.2611485

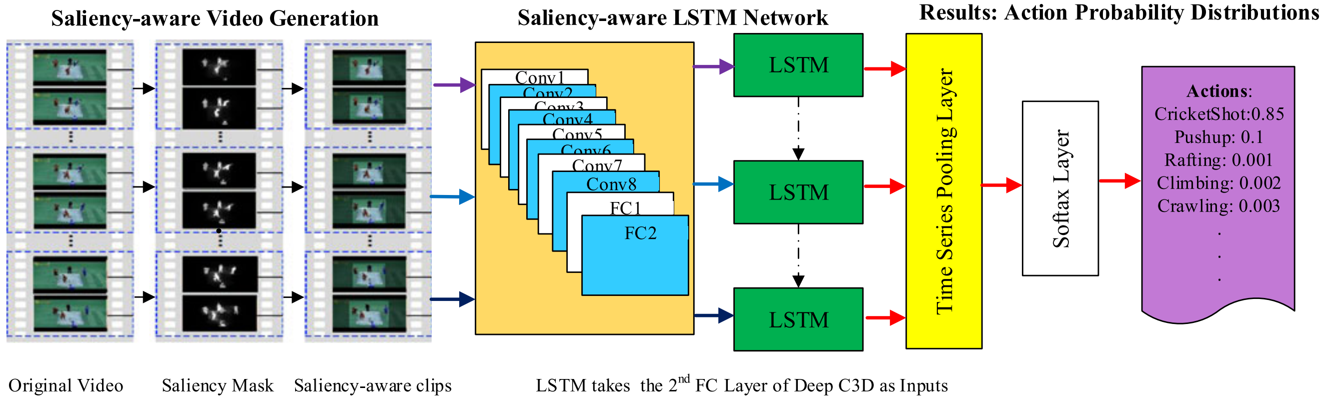

Fig. 1. The general framework of our saliency-aware 3-D CNN with LSTM for video activity recognition. scLSTM consists of two phases. First, we apply saliency-aware methods to generate salience-aware videos. Then, we design an end-to-end pipeline by integrating 3-D CNN with LSTM, followed by a time series pooling layer and a softmax layer to predict the activities of videos.

Recent efforts have shown that long short-term memory (LSTM) [2], [12] is able to learn when to forget previous hidden states and when to update hidden states by integrating memory units. A new record [2] was recently set on several benchmark datasets by feeding frame-level CNN sequence features to LSTM model for video activity recognition. These recurrent model based visual recognition pipeline are a natural choice for perceptual problems with time-varying visual input or sequential outputs. LSTM has been successfully adopted to several tasks, e.g., speech recognition [6], language translation [20], and image caption [22]. However, the above pipeline takes frame-level CNN sequence features as input for LSTM, which may fail to capture the rich motion information from adjacent frames. Furthermore, an activity is conducted by a subject or multiple subjects. It is important to consider attention which allows for salient features, instead of mapping an entire frame into a static representation.

In order to deal with above two issues, we propose a novel pipeline, called saliency-aware 3-D CNN and LSTM (scLSTM) for video action recognition by integrating the LSTM with salient action motion detection. It is worth highlighting the following contributions:

1) We propose an end-to-end pipeline by integrating LSTM with 3-D CNN for video action recognition. The 3-D CNN features on videos shots contains richer motion information than frame-level CNN features, and LSTM can explore the temporal relationship between video shots;
2) Saliency is further introduced to capture important subjects from video shots, which will improve the performance of 3-D CNN features.
3) Our method set a new record on two benchmark datasets, i.e., UCF101 with 13 320 videos and HMDB-51 with 6766 videos. It outperforms the counterparts by 3.8% and 3.2%, respectively.

## II. The Proposed Approach

In this section, we introduce our method scLSTM that consists of two phases (see Fig. 1). First, we apply saliency-aware methods to generate salience-aware videos. Then, we design an end-to-end pipeline by integrating 3-D CNN with LSTM, followed by a time series pooling layer and a softmax layer to predict the activities.

### *A. Salience-Aware Video Generation*

Image region segmentation has shown to benefit many specific visual tasks and applications, such as object detection and action recognition [13]. Recently, several methods have managed to generate considerable object proposals in every frame and transfer the task of object segmentation into an object region selection problem by utilizing both motion and appearance information to calculate the objectness scores. For example, [24] has proposed a saliency-aware based video object segmentation method, which performs better than the state-of-the-art methods. Inspired by the success of saliency methods, in this letter, we aim to integrate saliency technique into our framework. Given a video, we process the video frame by frame. First, we use method in [24] to generate a saliency-aware map $\mathbf{M}$ for each frame. Then, the corresponding frame saliency mask is computed by binarizing the $\mathbf{M}$. If $m_{i,j} < \text{mean}\,(\mathbf{M})$, then $m_{i,j} = 0$, otherwise $m_{i,j} = 1$. Next, we weaken the background regions to improve the importance of foreground objects (i.e., subject salience information). This is conducted by halving the RGB values of background regions where $m_{i,j} = 0$. As a result, a salience-aware video is generated and denoted as $\mathbf{V}$.

### *B. Saliency-Aware 3-D CNN With LSTM*

As mentioned in the previous section, a salience-aware video is represented as $\mathbf{V}$. In this section, we integrate the deep 3-D CNN with LSTM to analyze $\mathbf{V}$ for action recognition. Suppose $\mathbf{V}$ has $N$ frames, and we denote $\mathbf{V} = (v_1, v_2, \ldots, v_N)$. First, we divide the video $\mathbf{V}$ into $T$ splits and $\mathbf{V} = (\mathbf{v}_1^s, \mathbf{v}_2^s, \ldots, \mathbf{v}_t^s, \ldots, \mathbf{v}_T^s)$, where $\mathbf{v}_t^s$ is the $t$th split of the $\mathbf{V}$, $T = \frac{N}{K}$, and $K$ is the length of the split. Next, we encode each split with a 3-D CNN network, thus a sequence of video shots is generated as $\mathbf{X} = (\mathbf{x}_1, \ldots, \mathbf{x}_T)$. Finally, a sequence model takes $\mathbf{X}$ as input to recognize actions.

*1) 3-D Signals Encoding:* Given a video split $\mathbf{v}_t^s$, we propose to model the saliency-aware video at the level of the temporal features $\mathbf{X} = (\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_t, \ldots, \mathbf{x}_T)$ that are extracted by the encoder. Specifically, we propose to use a 3-D CNN, which has recently been demonstrated to capture well temporal dynamics in video splits [21]. In this letter, we use a 3-D CNN to build the higher level representations that preserve and summarize the local motion of a video clip. According to [21], a deep 3-D CNN network contains eight 3-D convolution layers, five 3-D max-pooling layers, and two fully connected layers. For simplicity, we refer 3-D convolution as $C\left(k, d, f, s_t, s_p\right)$ and pooling kernels as $P\left(d, f, s_t, s_p\right)$, where $k$ is the number of kernels, $d$ is temporal depth, $f$ is the spatial size, $s_t$ is the temporal stride, and $s_p$ is the spatial stride. Using shorthand notations, the 3-D CNN that is used in our method can be represented as: Conv(64,3,3,1,1), Pool(1,2,1,2), Conv(128,3,3,1,1), Pool(2,2,2,2), Conv(256,3,3,1,1), Conv(256,3,3,1,1), Pool(2,2,2,2), Conv(512,3,3,1,1), Conv(512,3,3,1,1), Pool(2,2,2,2), Conv(512,3,3,1,1), Conv(512,3,3,1,1), Pool(2,2,2,2), FC(4096), FC(4096), where FC$(n)$ is full connected layer with 4096 output units.

*2) Sequence-to-Sequence Model:* Recent advances in machine translation has shown that recurrent neural network, especially LSTM has potential to efficiently map sequences to sequences by incorporating memory units [2].

The main idea to handle the relationship between several video clips is to first encode the salience-aware video clips with 3-D CNN, one at a time, representing the video using a set of latent vector representations, and then decode from that representations to action names. Let us briefly introduce the basic LSTM unit, which consists of a single memory cell, an input activation function, and three gates (input $i_t$, forget $f_t$, and output $o_t$). $i_t$ allows incoming signal to alter the state of the memory cell or block it. $f_t$ controls what to be remembered and what to be forgotten by the cell and somehow can avoid the gradient from vanishing or exploding when back propagating through time. Finally, $o_t$ allows the state of the memory cell to have an effect on other neurons or prevent it. These additions to the single memory cell enable LSTM to capture extremely complex and long-term temporal dynamics and to overcome the vanishing gradients problems. Based on the LSTM unit, for an input $x_t$ at time step $t$, the LSTM computes a hidden/control state $h_t$ and a memory cell state $c_t$, which is an encoding of everything the cell has observed until time $t$:

$$\begin{aligned}
i_t &= \sigma\left(W_{xi}x_t + W_{hi}h_{t-1} + b_i\right)\\
f_t &= \sigma\left(W_{xf}x_t + W_{hf}h_{t-1} + b_f\right)\\
o_t &= \sigma\left(W_{xo}x_t + W_{ho}h_{t-1} + b_o\right)\\
g_t &= \sigma\left(W_{xg}x_t + W_{hg}h_{t-1} + b_g\right)\\
c_t &= f_t \odot c_{t-1} + i_t \odot g_t\\
h_t &= o_t \odot \phi\left(c_t\right)
\end{aligned}$$

where $\sigma$ is the sigmoidal nonlinearity, $\phi$ is the hyperbolic tangent non-linearity, $\odot$ is the element-wise product with the gate value, $W_{ij}$ is the weight matrices, and $b_j$ is the bias, $\sigma(x)$ is the logistic sigmoid nonlinearity, and $\phi(x)$ is the hyperbolic tangent activation function.

*3) LSTM With Time Series Pooling:* Temporal feature pooling can be incorporated directly as a layer and has been extensively used for video classification [12], [14]. This allows us to implement with the location of the temporal pooling layer with respect to the LSTM network architecture. By exploring various types of temporal pooling, we consider using both mean-pooling and max-pooling, which have several desirable properties shown in [12]. Next, we concatenate the mean-pooling and max-pooling features into a vector $\mathbf{Z}$ as the final video-level descriptor and feed it into a classifier loss layer.

Thus, given an input sequence $\mathbf{X} = \left(\mathbf{x}_1, \ldots, \mathbf{x}_T\right)$ in the encoding phase, the LSTM computes a sequence of hidden states $\left(h_1, h_2, \ldots, h_L\right)$. A probability distribution of an action category $P_{U,W}\left(y\right)$ is calculated by taking a softmax over the output $\mathbf{Z}$ of the time series pooling layer. $U$ is the parameters for the C3D network. Finally, The distribution is computed over a space $C$ (action categories) by the following equation:

$$\begin{aligned}
&P_{U,W}\left(y = c|\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_t, \ldots, \mathbf{x}_T\right)\\
&\quad = \frac{\exp\left(S_{zc}\mathbf{Z} + b_c\right)}{\sum_{c' \in C}\exp\left(S_{zc'}\mathbf{Z} + b_{c'}\right)}
\end{aligned}$$

where $S_{zc}, S_{zc'}$, $b_c$, and $b_{c'}$ are the parameters for the softmax layer.

TABLE I
RESULTS OF DIFFERENT SETTINGS ON THE UCF101 (SPLIT 1)

| Training setting | Accuracy(%) |
|---|---|
| 3-D CNN | 80.0 |
| 3-D CNN + LSTM | 82.8 |
| 3-D CNN + LSTM + Time_Pooling | 83.6 |
| Saliency + 3-D CNN +<br>LSTM + Time_Pooling | **84.7** |

TABLE II
ACTION RECOGNITION RESULTS ON UCF101 AND HMDB51 OVER 3 SPLITS: SPLIT@1, SPLIT@2, AND SPLIT@3

| Method | HMDB51 | UCF101 |
|---|---|---|
| iDT w/BoW + linear SVM [23] | 52.1 | 76.2 |
| Imagenet + linear SVM [21] | – | 68.8 |
| Deep 3-D CNN + linear SVM [21] | – | 82.3 |
| Deep networks [8] | – | 65.4 |
| LRCN [2] | – | 71.1 |
| Spatial Stream Network [16] | 40.5 | 73.0 |
| FstCN (only 1 path) [19] | 49.3 | 76.0 |
| Deep 3-D CNN [21] | 51.9 | 81.2 |
| **scLSTM** | **55.1** | **84.0** |

Our method is compared with baselines and current state-of-the-art methods. Top: The best handcrafted features with linear SVM. Middle: 2-D CNN and 3-D CNN features with linear SVM (none end-to-end). Bottom: End-to-end methods taking only RGB frames as inputs.

### C. Optimization

In our framework, the weight parameters of scLSTM can be learned jointly by using mini-batch SGD algorithm to minimize the negative log likelihood $L\left(U, W, S\right) = \sum_{i \in D} -\log\left(P_{U,W}\left(y\right)\right)$. The batch size is set to 30. Learning rate is set to $1\mathrm{e}^{-4}$. The optimization is finished after 80k iterations. To mitigate the risk of overfitting, we use the pretrained C3D model trained on the large-scale Sports-1M dataset in [21] to initialize 3-D CNN and then fine tune it on the target dataset. One of the most appealing aspects of the described approach is the ability to learn the parameters “end-to-end”.

## III. EXPERIMENTS

We conduct experiments with two goals. First, we study the influence of LSTM and saliency in our algorithm. Second, we compare our results with other state-of-the-art algorithms.

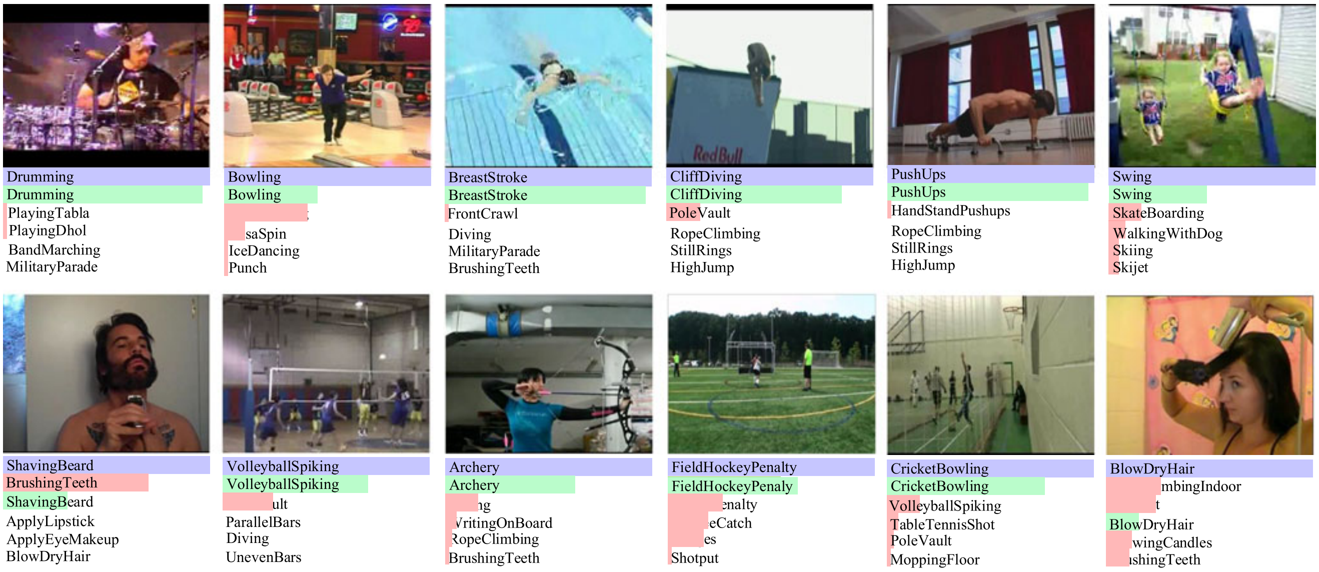

Fig. 2. Predictions on UCF-101 test data. Purple (first row) indicates ground truth label and the bars below show model predictions sorted in decreasing confidence. Green and red distinguish correct and incorrect predictions, respectively.

### *A. Datasets*

In this experiment, we use two datasets: UCF101 dataset [18] and HMDB-51 [10]. Specifically, UCF101 consists of 13 320 videos of 101 human action categories. It is one of two most challenging datasets to date. Moreover, the HMDB-51 dataset has 6766 videos organized as 51 distinct action categories. They are collected from a wide range of sources. This dataset is quite challenging for it contains complex context environments and a small amount of training videos. In addition, both UCF-101 and HMDB-51 have three split settings, thus we report the mean accuracy over three splits.

### *B. Baselines and Evaluation Metrics*

We compare our method with a few baselines:

1) The current best hand-crafted feature: iDT that normalize histogram of trajectories, HOG, HOG, and MBH features to form a 25 000-dimensional feature vector for a video.
2) Deep 2-D CNN and 3-D CNN features are separately extracted and then input to a linear SVM.
3) End-to-end methods taking only RGB frames as inputs, such as deep networks [8], spatial stream network [16], long-term recurrent convolutional network (LRCN) [2], and factorized spatio-temporal convolutional networks (FstCN) [19].

We report the mean accuracy for each of these methods.

### *C. Comparison of Network Architectures*

We first study the effect of the submodule of the scLSTM framework and compare different architectures on the split 1 of the UCF101 dataset. In experiments, we fix the value of $K$ at 16 and the max number of T at 10. Results in Table I show that 3-D CNN + LSTM outperforms 3-D CNN. The increased accuracy is probably due to the advances in considering the relationship between video clips, which is also depicted in LRCN [2]. From the results, we find that time series pooling over the output of LSTM provides better performance than LSTM without time series pooling. Moreover, integrating salience-aware mechanism with time series pooling based LSTM provides the best performance 84.7%. This suggests that the performance of action recognition largely depends on the relationship between clips and the salience regions.

### *D. Results*

We show that the comparison of our approach with the baselines in Table II and Fig. 2. First, the experimental results on HMDB51 show that our approach outperforms the state-of-the-art methods. Specifically, our proposed method outperforms the currently best approach (deep 3-D CNN) on mean accuracy by 3.2%. Second, we have several observations from results on UCF101.

1) iDT performs better than Imagenet, but the dimension of the iDT (i.e, 25k) is extremely higher than Imagenet (i.e., 4096).
2) Deep 3-D CNN performs better than iDT features and 2-D CNN features.
3) Based on the only RGB and end-to-end setting, our approach outperforms the best state-of-the-art methods by 3.8%.

Third, the quantitative examples in Fig. 2 shows that our approach can achieve promising performance in practice.

## IV. Conclusion

In this letter, we propose a general framework for video action recognition. We first utilize saliency-aware method to generate a video that enhances the importance of foreground regions. Next, we integrate C3D net, LSTM, and time pooling for extracting the most representative features for videos. Experiments on the UCF101 and HMDB51 demonstrate the superiority of our scLSTM compared to others.

## References

[1] P. Dollar, V. Rabaud, G. Cottrell, and S. Belongie, "Behavior recognition via sparse spatio-temporal features," in *Proc. 14th Int. Conf. Comput. Commun. Netw.*, 2005, pp. 65–72.

[2] J. Donahue *et al.*, "Long-term recurrent convolutional networks for visual recognition and description," in *Proc. 2015 Comput. Vis. Pattern Recognit.*, 2015.

[3] L. Gao, J. Song, X. Liu, J. Shao, J. Liu, and J. Shao, "Learning in high-dimensional multimedia data: the state of the art," *Multimedia Syst.*, pp. 1–11, 2015.

[4] L. Gao, J. Song, F. Nie, Y. Yan, N. Sebe, and H. T. Shen, "Optimal graph learning with partial tags and multiple features for image and video annotation," in *Proc. 2015 IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2015, pp. 4371–4379.

[5] L. Gao, J. Song, F. Nie, F. Zou, N. Sebe, and H. T. Shen, "Graph-without-cut: An ideal graph learning for image segmentation," in *Proc. 13th AAAI Conf. Artif. Intell., Phoenix, Arizona, USA.*, 2016, pp. 1188–1194.

[6] A. Graves and N. Jaitly, "Towards end-to-end speech recognition with recurrent neural networks," *Proc. 31st Int. Conf. Mach. Learn.* JMLR Workshop Conf. Proc., 2014, pp. 1764–1772.

[7] S. Ji, W. Xu, M. Yang, and K. Yu, "3d convolutional neural networks for human action recognition," *IEEE Trans. Pattern Anal. Mac. Intell.*, vol. 35, no. 1, pp. 221–231, Jan. 2013.

[8] A. Karpathy, G. Toderici, S. Shetty, T. Leung, R. Sukthankar, and F. Li, "Large-scale video classification with convolutional neural networks," in *Proc. 2014 Comput. Vis. Pattern Recognit.*, 2014, pp. 1725–1732.

[9] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," in*Advances in Neural Information Processing Systems 25*, F. Pereira, C. J. C. Burges, L. Bottou, and K. Q. Weinberger, Eds.. New York, NY, USA: Curran Associates, 2012, pp. 1097–1105.

[10] H. Kuehne, H. Jhuang, E. Garrote, T. Poggio, and T. Serre, "HMDB: A large video database for human motion recognition," in *Proc. Int. Conf. Comput. Vis.*, 2011, pp. 2556–2563.

[11] I. Laptev, M. Marszalek, C. Schmid, and B. Rozenfeld, "Learning realistic human actions from movies," in *Proc. 2008 Comput. Vis. Pattern Recognit.*, 2008.

[12] J. Y.-H. Ng, M. Hausknecht, S. Vijayanarasimhan, O. Vinyals, R. Monga, and G. Toderici, "Beyond short snippets: Deep networks for video classification," in *Proc. 2015 IEEE Conf., Comput. Vis. Pattern Recog.*, Jun. 2015, pp. 4694–4702.

[13] S. Ren, K. He, R. Girshick, and J. Sun, "Faster R-CNN: Towards real-time object detection with region proposal networks," in *Advances Neural Inform. Process. Syst.*, 2015.

[14] M. S. Ryoo, B. Rothrock, and L. Matthies, "Pooled motion features for first-person videos," in *Proc. Comput. Vis. Pattern Recognit.*, 2015.

[15] P. Scovanner, S. Ali, and M. Shah, "A 3-dimensional sift descriptor and its application to action recognition," in *Proc. ACM Multimedia*, pp. 357–360, 2007.

[16] K. Simonyan and A. Zisserman, "Two-stream convolutional networks for action recognition in videos," in *Proc. Neural Inform. Process. Syst.*, pp. 568–576, 2014.

[17] J. Song, L. Gao, F. Nie, H. Shen, Y. Yan, and N. Sebe, "Optimized graph learning with partial tags and multiple features for image and video annotation," *IEEE Trans. Image Process.*, 2016.

[18] K. Soomro, A. R. Zamir, and M. Shah, "UCF101: A dataset of 101 human actions classes from videos in the wild," *CoRR*, arXiv:1212.0402, 2012.

[19] L. Sun, K. Jia, D. Yeung, and B. E. Shi, "Human action recognition using factorized spatio-temporal convolutional networks," *CoRR*, arXiv:1510.00562, 2015.

[20] I. Sutskever, O. Vinyals, and Q. V. Le, "Sequence to sequence learning with neural networks," in *Advances Neural Information Processing Systems 27*, Z. Ghahramani, M. Welling, C. Cortes, N. D. Lawrence, and K. Q. Weinberger, Eds. New York, NY, USA: Curran Associates, 2014, pp. 3104–3112.

[21] D. Tran, L. Bourdev, R. Fergus, L. Torresani, and M. Paluri, "Learning spatiotemporal features with 3d convolutional networks," in *Proc. 2015 IEEE Int. Conf. Comput. Vis.*, Dec. 2015, pp. 4489–4497.

[22] O. Vinyals, A. Toshev, S. Bengio, and D. Erhan, "Show and tell: A neural image caption generator," in *Proc. IEEE Conf. Comput. Vis. Pattern Recog.*, Jun. 2015, pp. 3156–3164.

[23] H. Wang and C. Schmid, "Action recognition with improved trajectories," in *Proc. IEEE Int. Conf. Comput. Vis.*, Sydney, Australia, 2013, pp. 3551–3558.

[24] W. Wang, J. Shen, and F. Porikli, "Saliency-aware geodesic video object segmentation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recog.*, Jun. 2015, pp. 3395–3402.

[25] M. D. Zeiler and R. Fergus, "Visualizing and understanding convolutional networks," in *Proc. 13th Eur. Conf. Comput. Vis.*, 2014, pp. 818–833.

[26] S. Zhang, D. Cheng, M. Zong, and L. Gao, "Self-representation nearest neighbor search for classification," *Neurocomputing*, vol. 195, pp. 137–142, 2016.

[27] X. Zhu, X. Li, and S. Zhang, "Block-row sparse multiview multilabel learning for image classification," *IEEE Trans. Cybern.*, vol. 46, no. 2, pp. 450–461, Feb. 2016.

[28] X. Zhu, X. Li, S. Zhang, C. Ju, and X. Wu, "Robust joint graph sparse coding for unsupervised spectral feature selection," *IEEE Trans. Neural Netw. Learning Syst.*, 2016.

[29] X. Zhu, L. Zhang, and Z. Huang, "A sparse embedding and least variance encoding approach to hashing," *IEEE Trans. Image Process.*, vol. 23, no. 9, pp. 3737–3750, Sep. 2014.

[30] X. Zhu, S. Zhang, Z. Jin, Z. Zhang, and Z. Xu, "Missing value estimation for mixed-attribute data sets," *IEEE Trans. Knowl. Data Eng.*, vol. 23, no. 1, pp. 110–121, Jan. 2011.

[31] Q. Liu, W. Cai, J. Shen, Z. Fu, X. Liu, and N. Linge, "A speculative approach to spatial-temporal efficiency with multi-objective optimization in a heterogeneous cloud environment," *Secur. Commun. Netw.*, vol. 9, no. 17, pp. 4002–4012, Nov. 2016.

[32] Z. Pan, J. Lei, Y. Lei, Y. Zhang, X. Sun, and S. Kwong, "Fast motion estimation based on content property for low-complexity H.265/HEVC encoder," *IEEE Trans. Broadcasting*, vol. 62, no. 3, pp. 675–684, Sep. 2016.

[33] Z. Pan, Y. Zhang, and S. Kwong, "Efficient motion and disparity estimation optimization for low complexity multiview video coding," *IEEE Trans. Broadcasting*, vol. 61, no. 2, pp. 166–176, Jun. 2015.